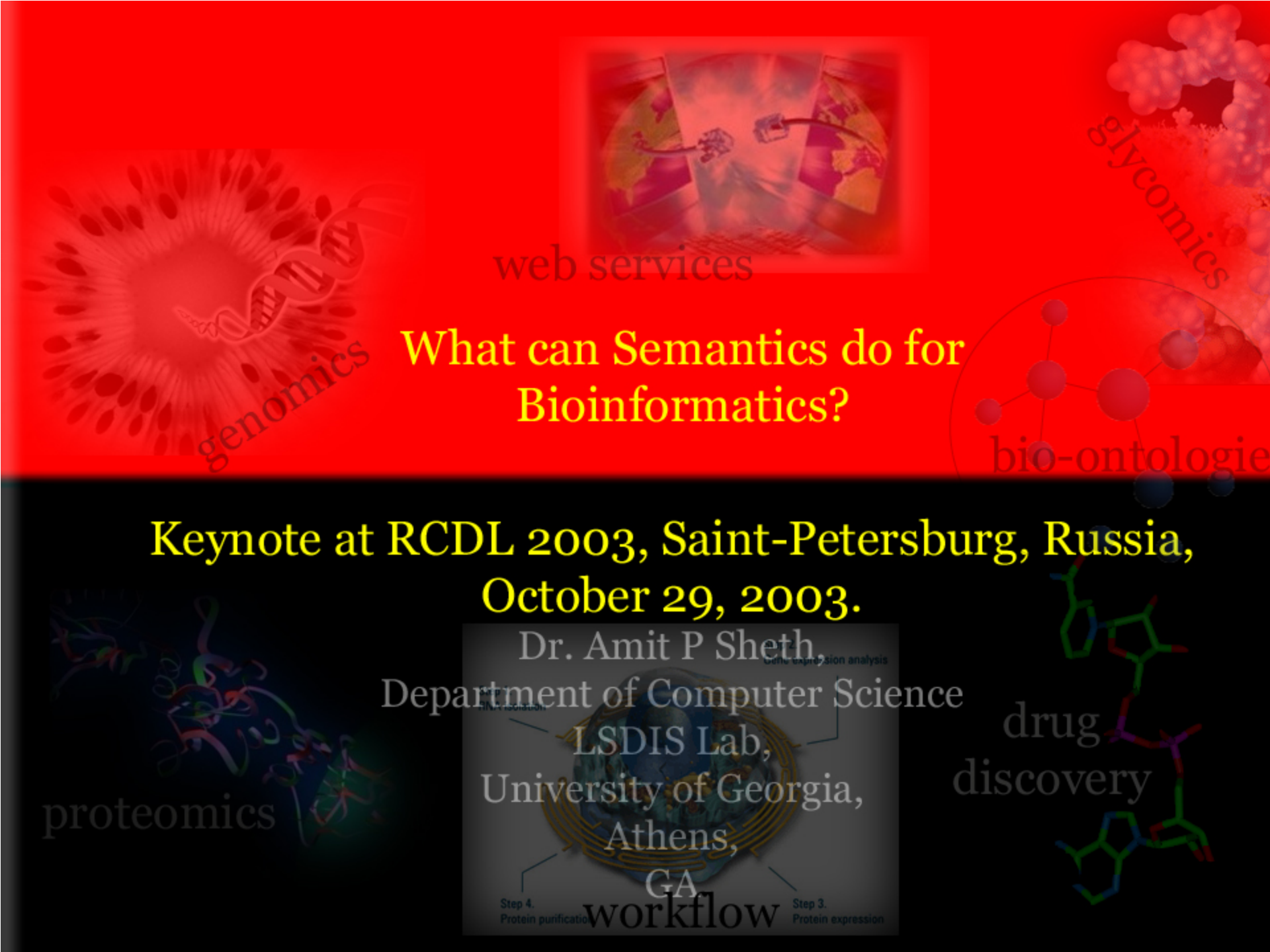
10-29-2003

What Can Semantics do for Bioinformatics?

Amit P. Sheth
Wright State University - Main Campus, amit@sc.edu



What can Semantics do for Bioinformatics?

Keynote at RCDL 2003, Saint-Petersburg, Russia, October 29, 2003.

Dr. Amit P Sheth,
Department of Computer Science
LSDIS Lab,
University of Georgia,
Athens,
GA

bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

Acknowledgements

UGA Biologists/Biochemists

- Will York, Complex Carbohydrate Research Center
- Jonathan Arnold, Fungal Genomics
- Phillip Bowen, CCQC

Project members of LSDIS lab projects
Bioinformatics for Glycan Expression and METEOR-S (incl. Miller, Kochut, Arpinar)

Special thanks in background research & preparation:
Karthik Gomadam, Christopher Thomas, Kunal Verma

Excellent starting point for complementary material
(a partial list)

- "Building a Bioinformatics Nation," Lincoln Stein's Keynote at O'Reilly's Bioinformatics Technology Conference 2002
- "Bio-Ontologies: Their creation and design" Peter Karp, Robert Stevens and Carole Goble
- "Query Processing with Description Logic Ontologies over Object-Wrapped Databases" Martin Peim, Enrico Franconi, Norman Paton and Carole Goble
- "Ontologies for molecular biology and bioinformatics" Steffen Schulze-Kremer (paper)
- "Can we do better than Google? Using semantics to explore large heterogeneous knowledge sources," Anatole Gershman, SWDB Workshop, 2003.

bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

Some current BioInformatics Systems

- Tambis
- BioMediator
- Biodas
- BioSem

Data integration, in some cases using an ontology. Single access point for multiple biological information sources; querying multiple sources.

bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

Outline of this talk...

- A Short History of Science
- Challenges in biology
- What can BioInformatics do for Biology?
- What can Semantics do for BioInformatics?
  - Some examples of Semantics-powered Bioinformatics

bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

What is difficult, tedious and time consuming now ...

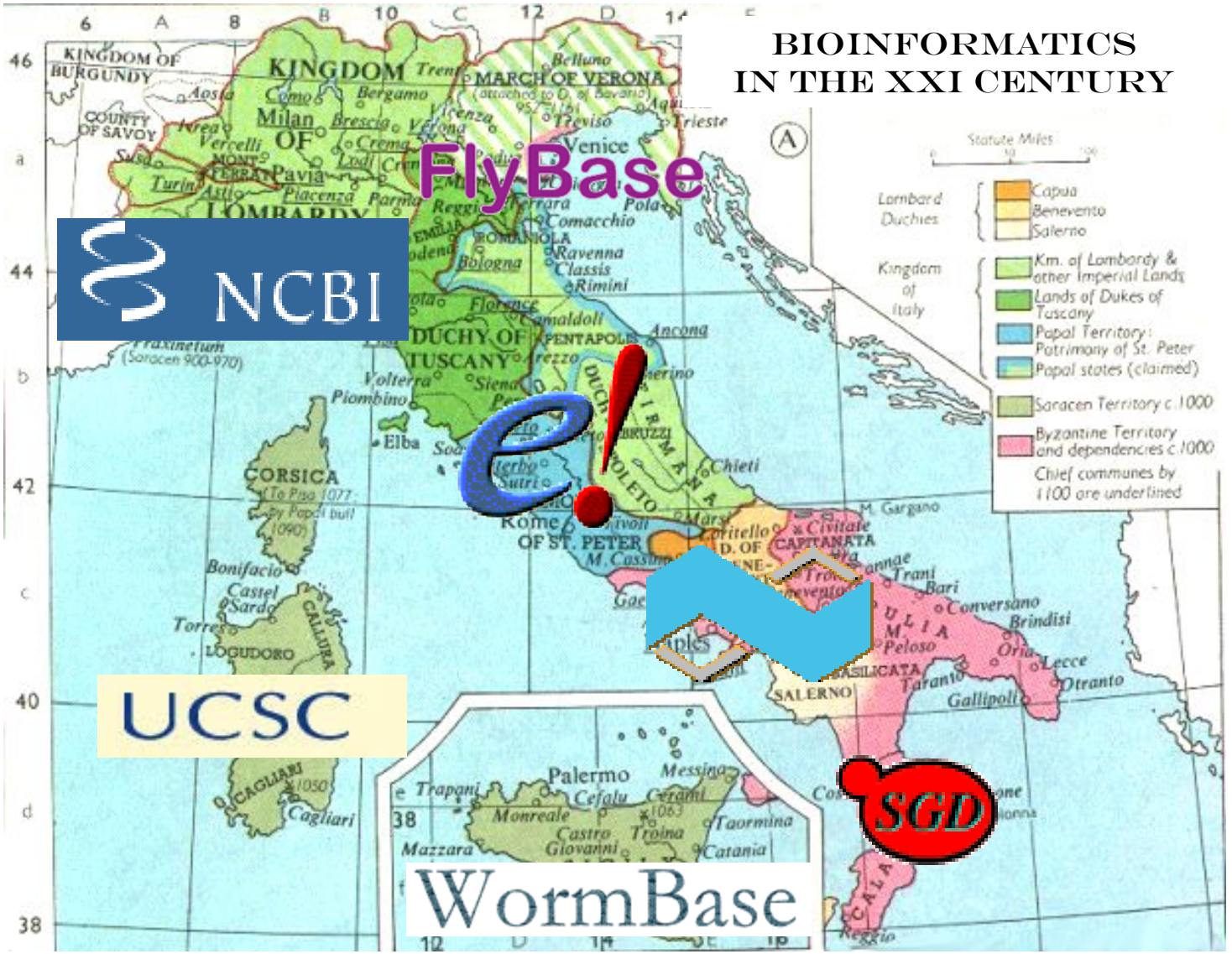
What genes do we all have in common?*

Research to answer this question took scientists two years**

* G. Strobel and J Arnold. Essential Eukaryotic Core, Evolution (to appear, 2003)

**but we now believe with semantic techniques and technology, we can answer similar questions much faster

Why? Bioinformatics, ca. 2002



From http://prometheus.frii.com/~gnat/tmp/stein_keynote.ppt

Science then, then and now

In the beginning, there was thought and observation.

Science then, then and now

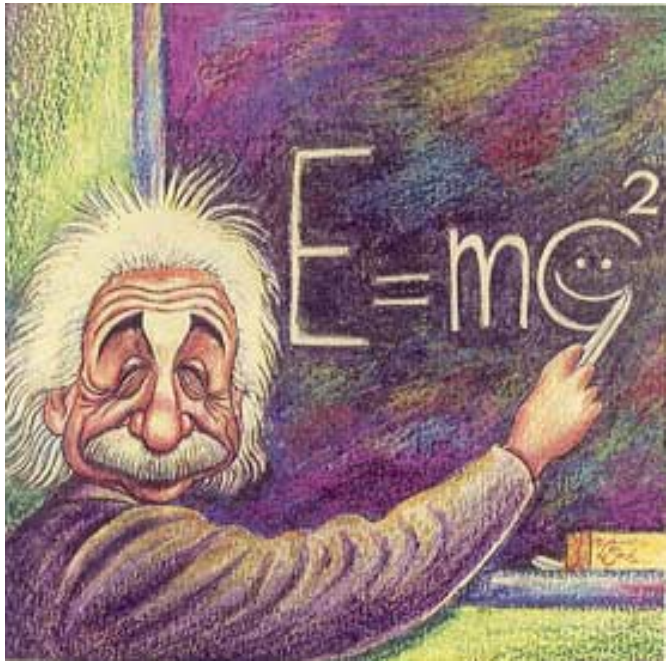
For a long time this didn't change.

- Man thought it would be enough to reason about the existing knowledge to explore everything there is to know.

- Back then, one single person could possess all knowledge in his cultural context.

The achievements are still admirable …

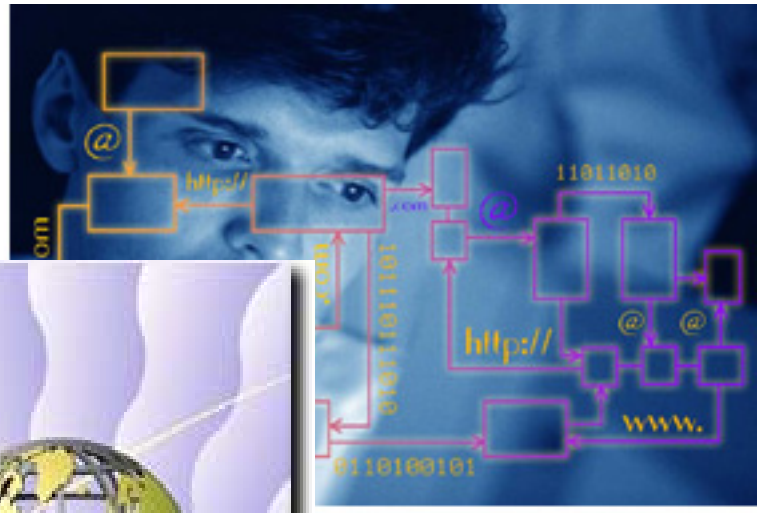
…as we can see



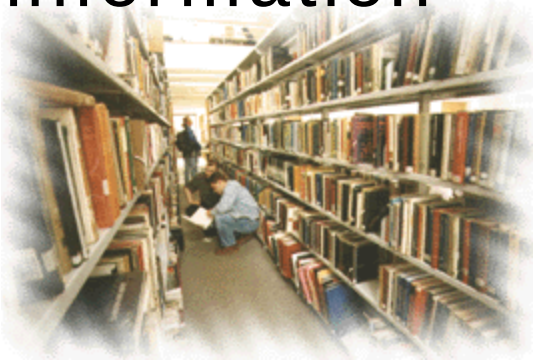
Reasoning and mostly passive observation were the main techniques in scientific research until recently.

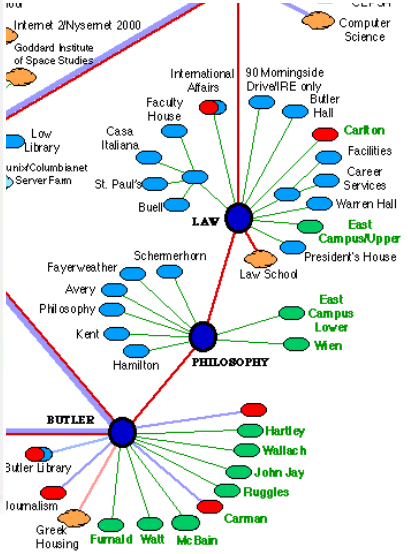
bio ontologies
bioinformatics
drug discovery
semantic web process
glycomics
LSDIS

Science then, then and now



A vast amount of information





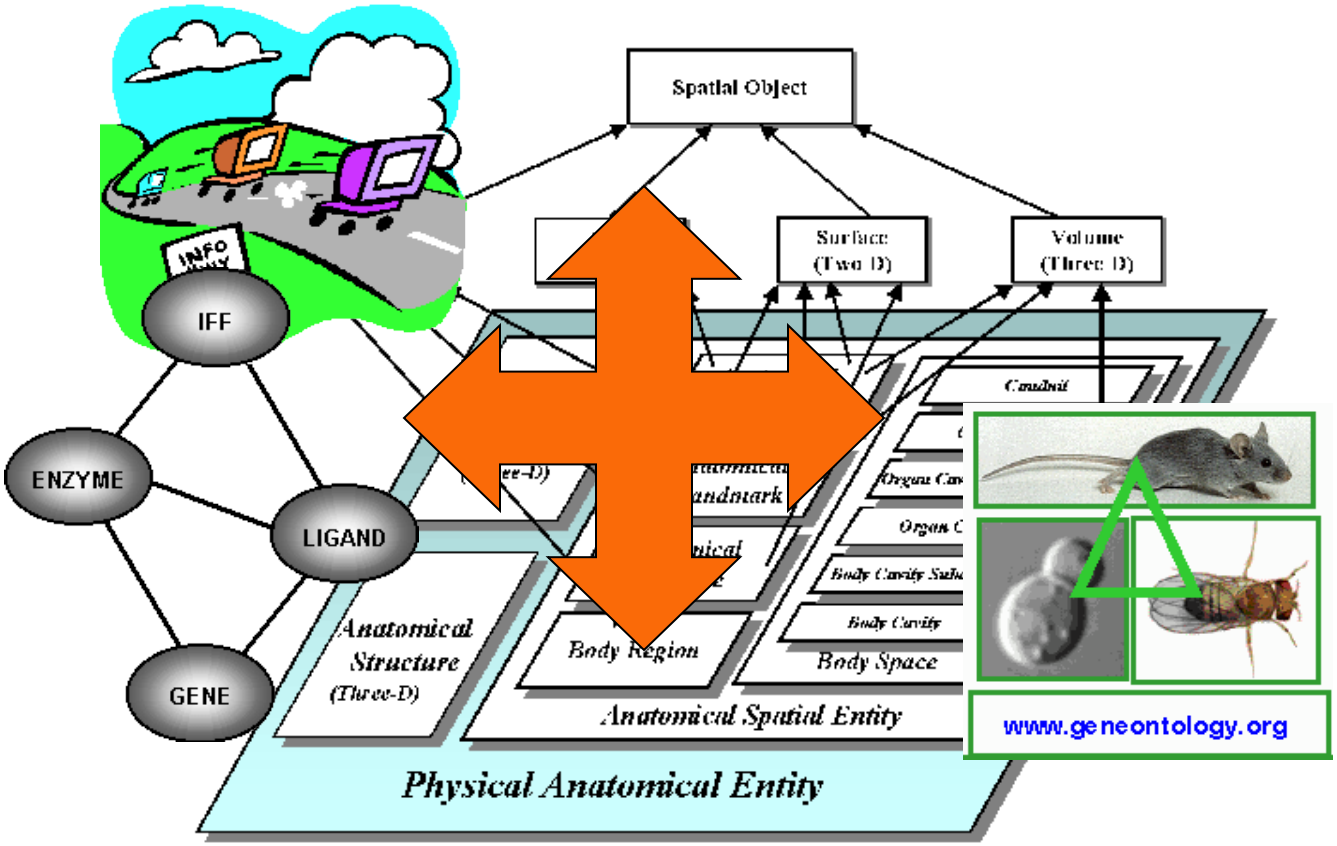
Science then, then and now

No single person, no group has an overview of what is known.

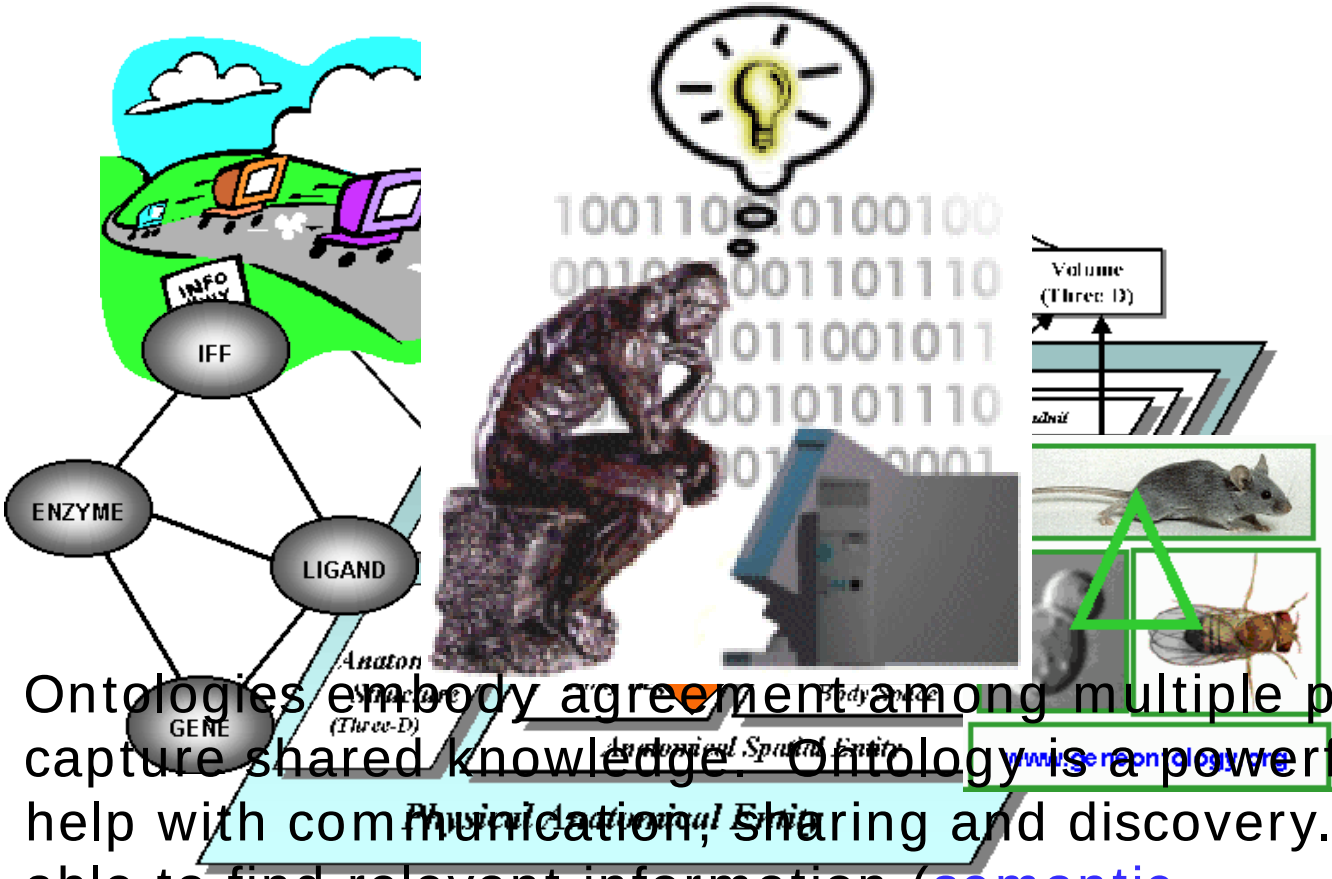
Known, But not known …→ not known

bio ontologies bioinformatics semantic web process LSDIS drug discovery glycomics

Science then, then and now



Science then, then and now



Ontologies embody agreement among multiple parties and capture shared knowledge. Ontology is a powerful tool to help with communication, sharing and discovery. We are able to find relevant information (semantic search/browsing), connect knowledge and information (semantic normalization/integration), find relationships between pieces of knowledge from different fields (gain insight, discover knowledge)

bio ontologies

bioinformatics

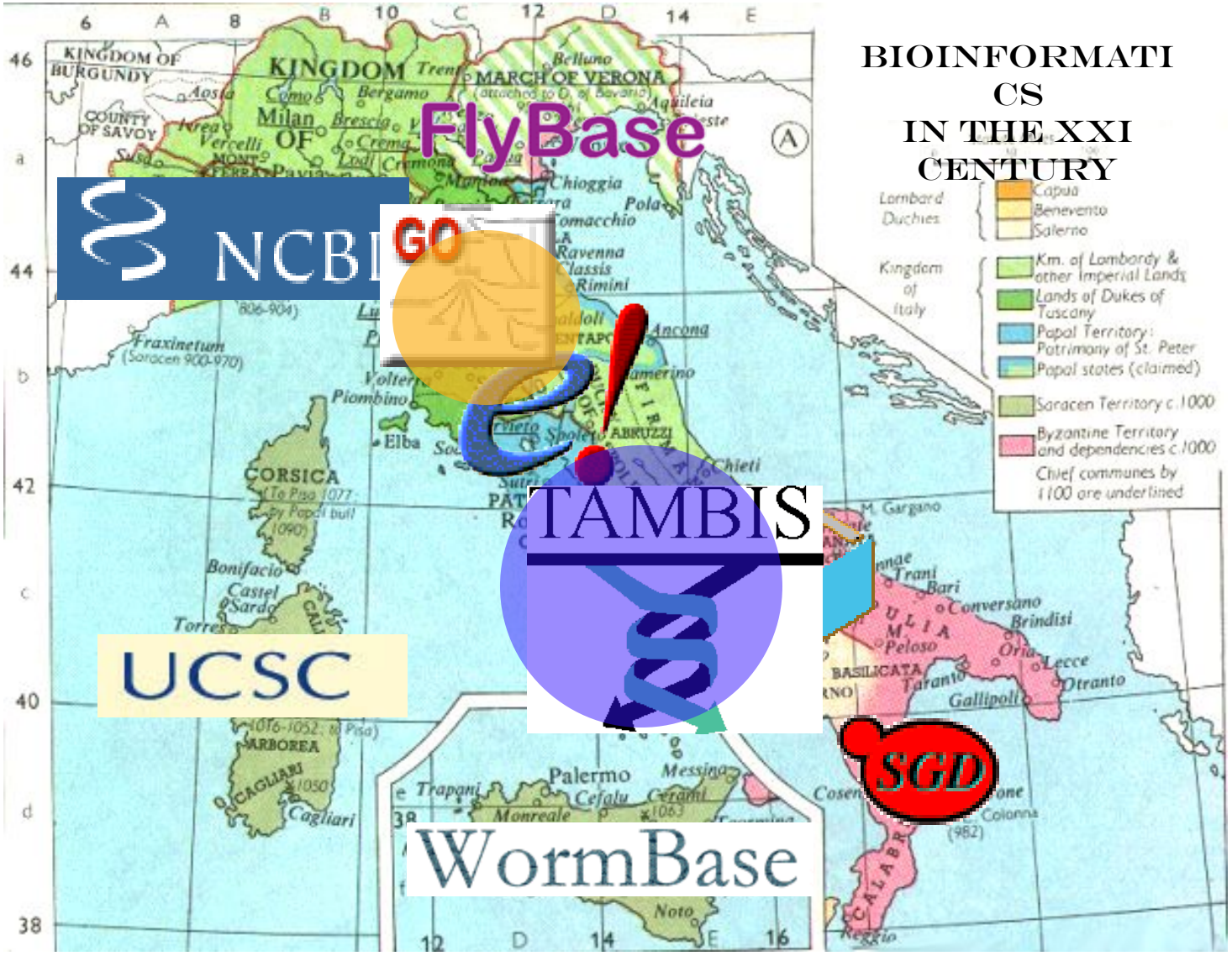
drug discovery

semantic web process

glycomics

LSDIS

Intervention by Ontologies …



bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

Outline of the talk…

- A Short History of Science
- Challenges in biology
- What can BioInformatics do for Biology?
- What can Semantics do for BioInformatics:
  - Some examples of Semantics-powered Bioinformatics

bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

Challenges in biology

- What makes us ill or unwell?
  - Disease identification, disease inducing agents
- What keeps us healthy and makes us live longer?
  - Drug discovery
- Where do we all come from and what are we made of?
  - Genetics and beyond

bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

… and their implications

- Understand biological structures of increasing complexity:
  - Genes (*Genomics*): 1980s
  - Proteins (*Proteomics*): 1990s
  - Complex Carbohydrates (*Glycomics*): 2000s
- Understand biological processes and the roles structures play in them (biosynthesis and biological processes)

Outline

- Evolution of Science
- Challenges in biology
- What can bioInformatics do for Biology?
- What can Semantics do for bioInformatics?
  - Some examples of Semantics-powered Bioinformatics

bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

What can BioInformatics do?

- Analyze genetic and molecular sequences
  - Look for patterns, similarities, matches
  - Identify structures
- Store derived information
  - Large databases of genetic information

bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

Outline

- Evolution of Science
- Challenges in biology
- What can bioInformatics do for Biology?
- What can Semantics do for bioInformatics?
  - Some examples of Semantics-powered Bioinformatics

bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

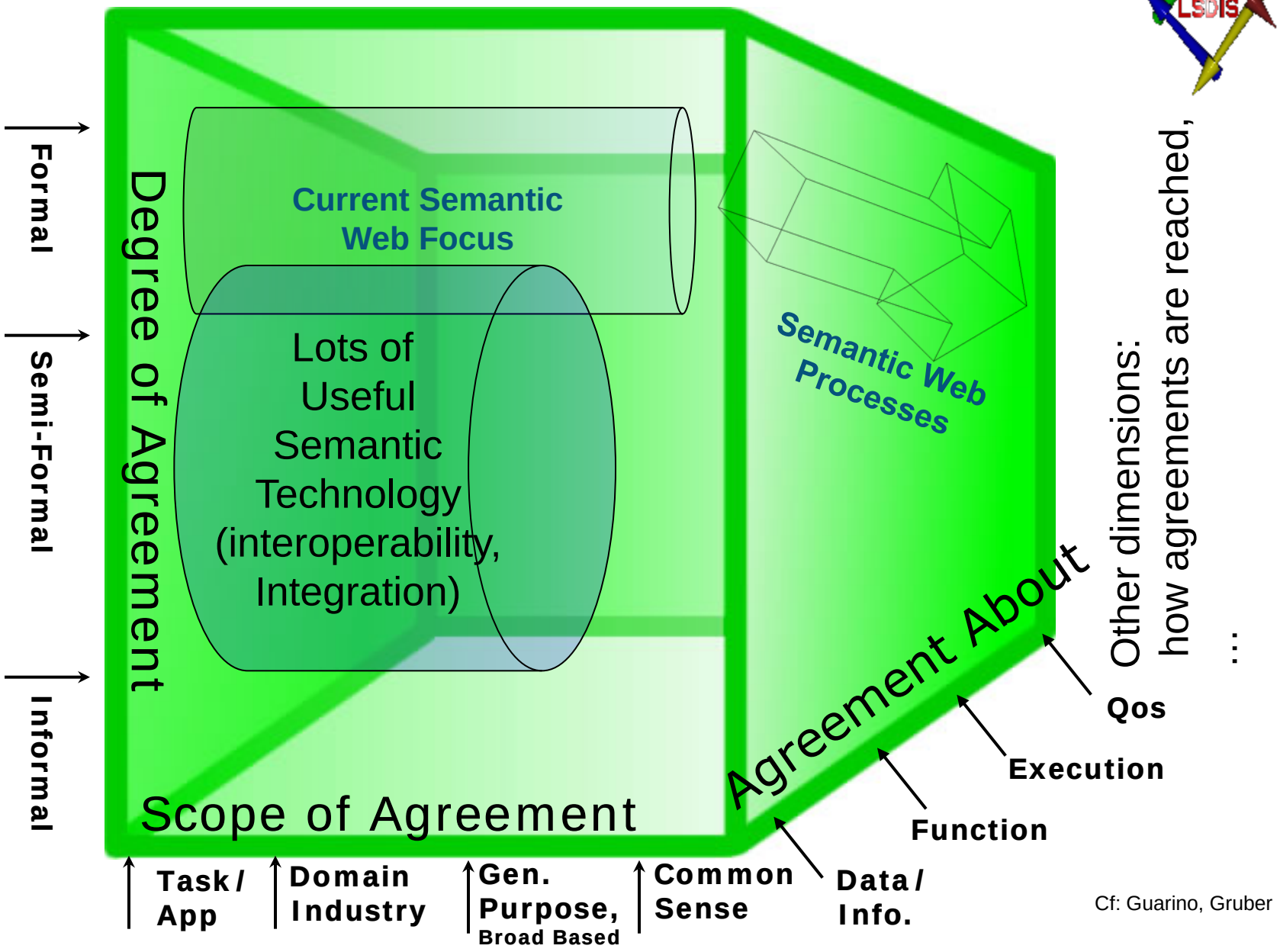
Paradigm shift over time: Syntax -> Semantics

Increasing sophistication in applying semantics & value add

- **Relevant Information (Semantic Search & Browsing)**
- **Semantic Information Interoperability and Integration**
- **Semantic Correlation / Association, Analysis, Insight and Discovery**

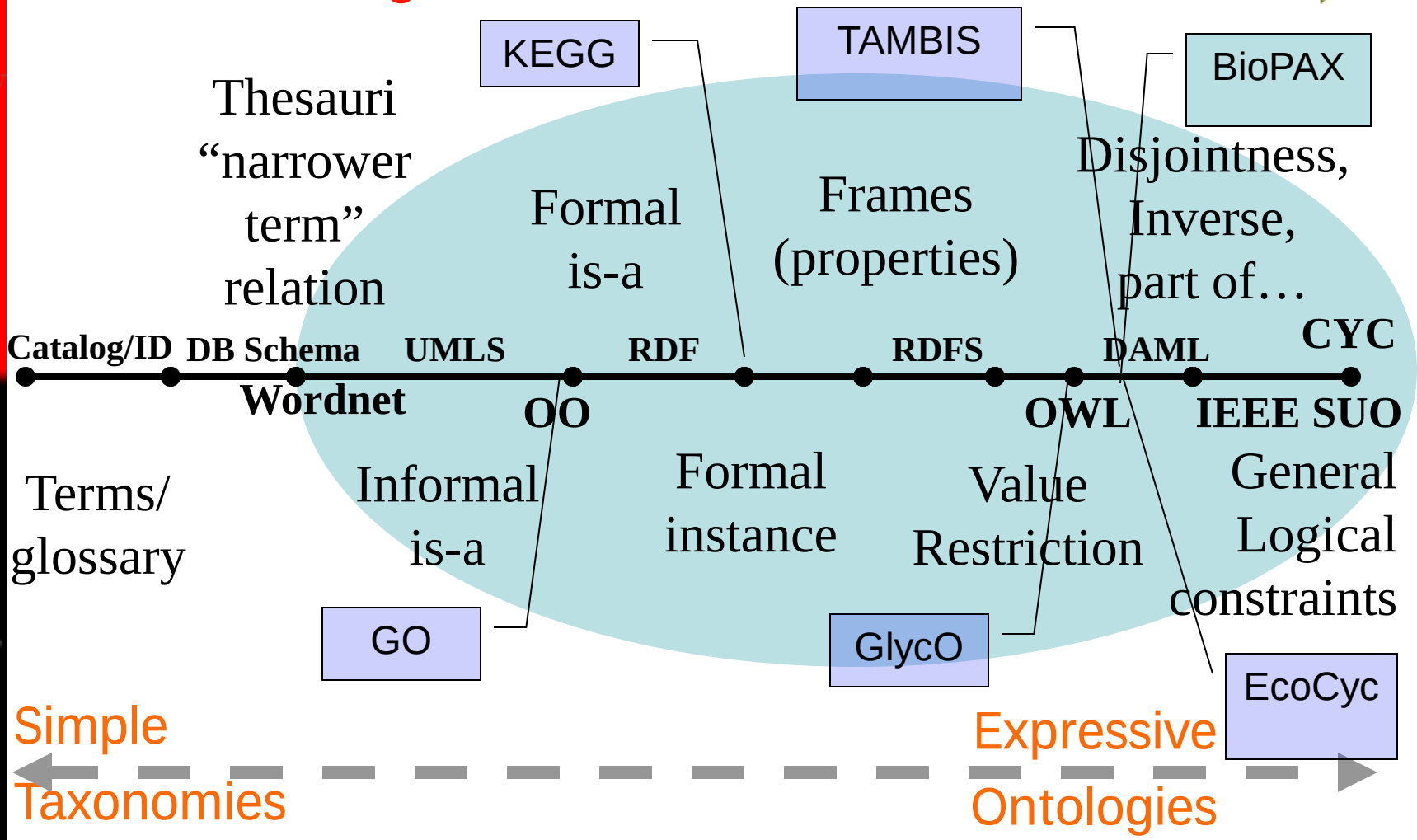
bio ontologies
bioinformatics
drug discovery
LSDIS
semantic web process
glycomics

Broad Scope of Semantic (Web) Technology



Cf: Guarino, Gruber

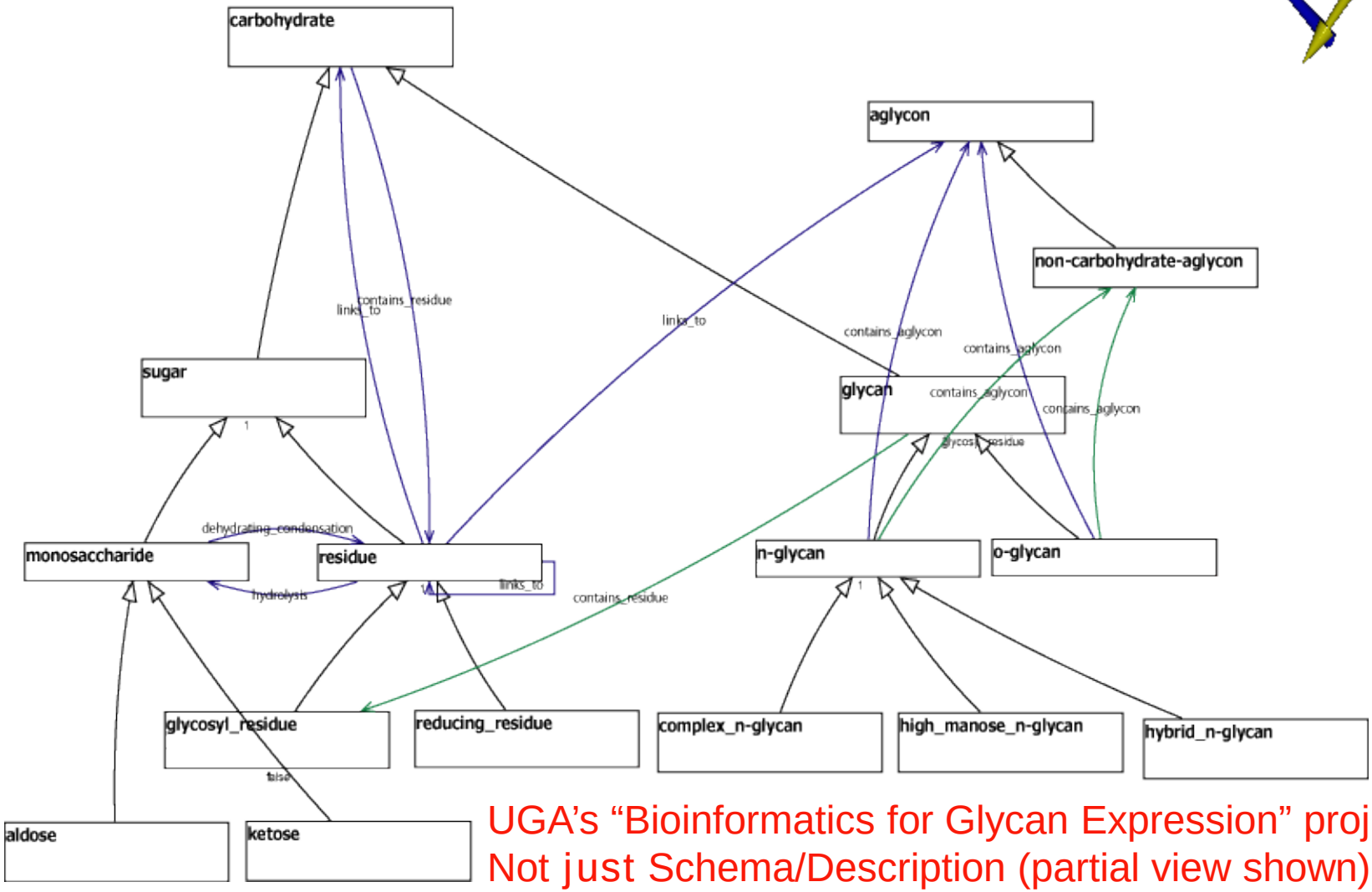
Knowledge Representation and Ontologies



Ontology Dimensions After McGuinness and Finin

bio ontologies bioinformatics glycomics drug discovery semantic web process LSDIS

GlycO: Glycan Structure Ontology



UGA's "Bioinformatics for Glycan Expression" proj.
Not just Schema/Description (partial view shown), also description base/ontology population.
In progress, uses **OWL.**

bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

What can current semantic technology do? (sample)

- Semi-automated (mostly automated) annotation of resources of various, heterogeneous sources (unstructured%, semi-structured, structured data; media content)*
- Creation of large knowledge bases (ontology population) from the trusted sources *
- Unified access to multiple sources*,#
- Inferenceing #
- Relationship/knowledge discovery among the annotated resources and entities; analytics*%
  - Both implicit^ and explicit* relationships

* Commercial: Semagix; %: Near-commercial: IBM/SemTAP;
# Commercial: Network Inference; ^ LSDIS-UGA Research

Industry Efforts

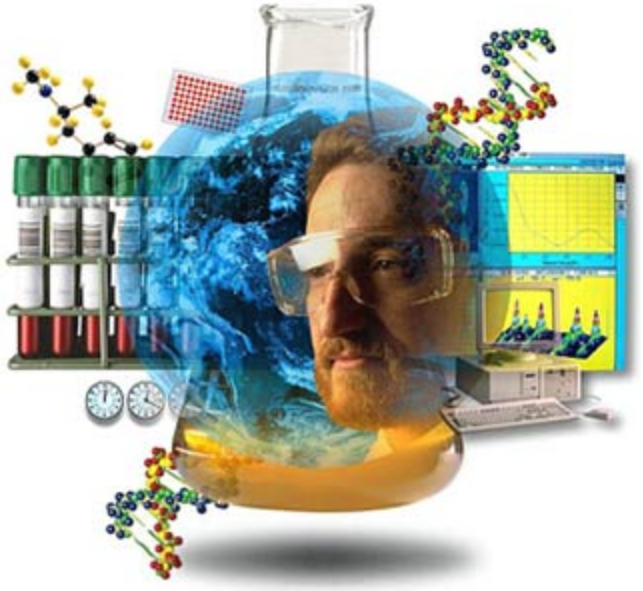
(examples with bioinformatics applications only)

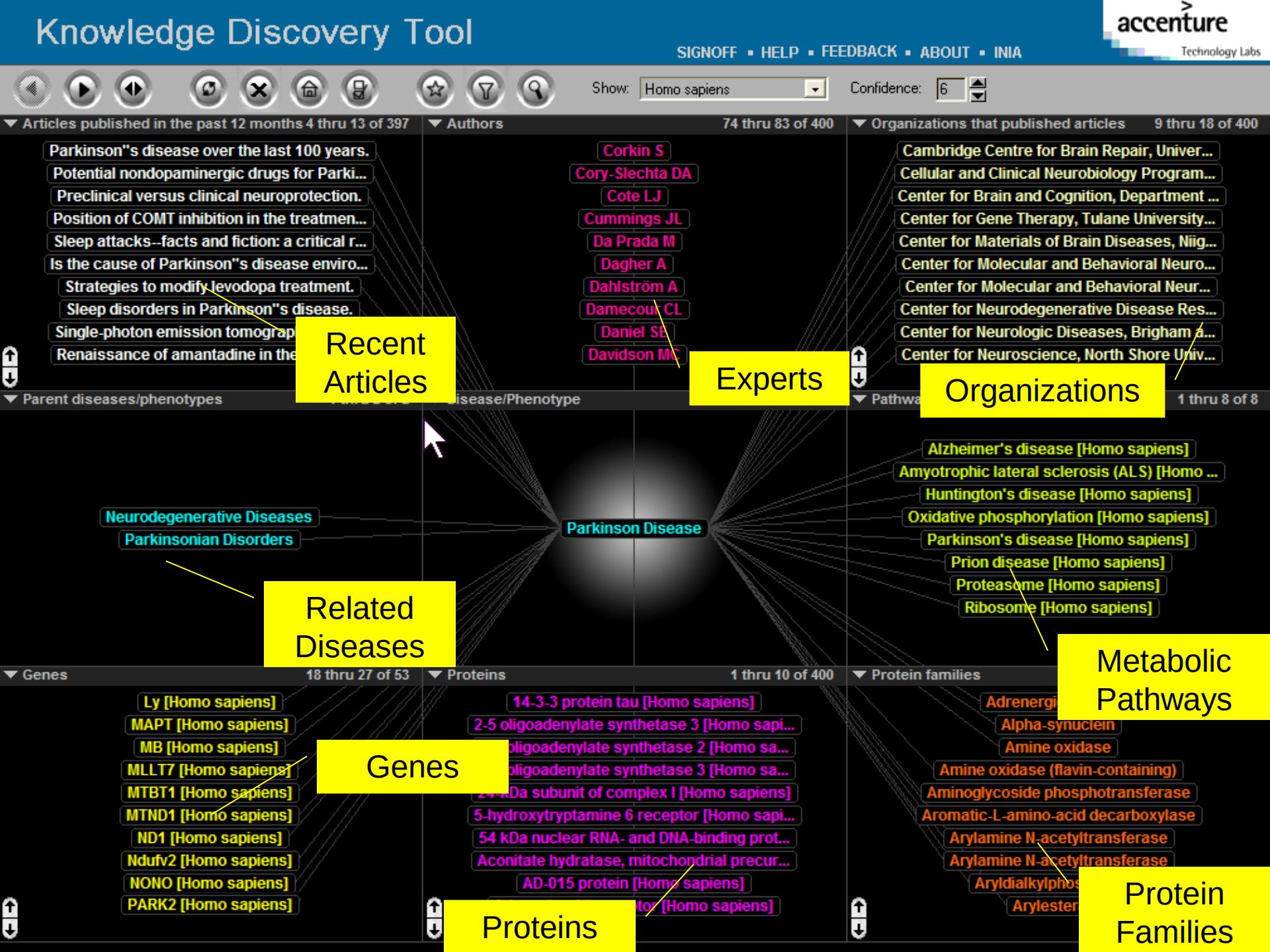
- Accenture's Knowledge Discovery Tool (pre-ontology?, not product)
- Semagix's Semantic Browsing and Querying application for drugs for Rogers MIS and its pharmaceutical customers (product and applications); also semantic analysis application (not discussed here)
- Network Inference's Cerebra server: semantic engineering based bioinformatics system aiding drug discovery (product?)

bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

Existing Systems using Semantics for Bioinformatics

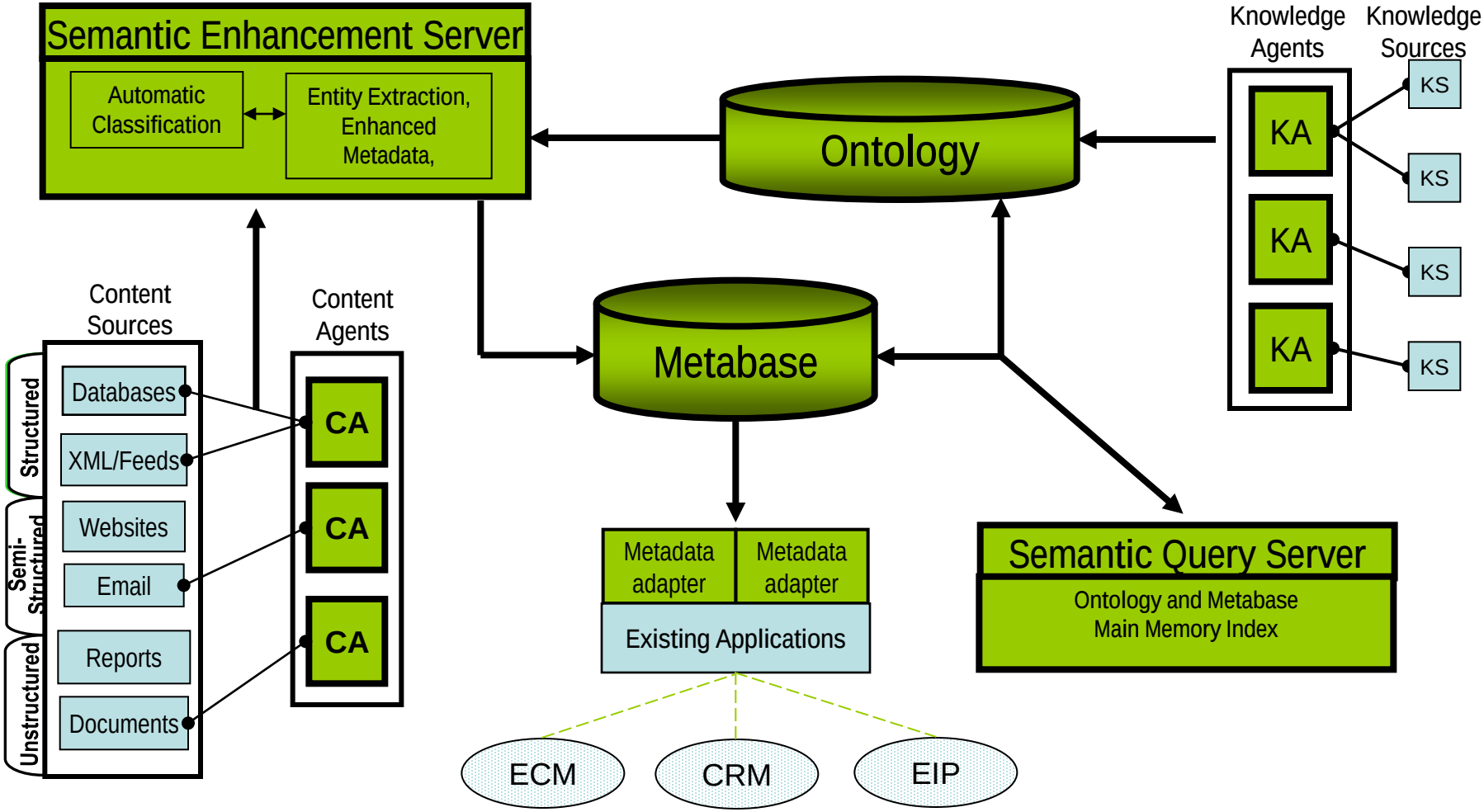
FOCUS: SEMANTIC SEARCH AND BROWSING (with nascent work in discovery)




Knowledge Discovery Tool
accenture
Technology Labs
SIGNOFF ▪ HELP ▪ FEEDBACK ▪ ABOUT ▪ INIA
Show: Homo sapiens
Confidence: 6
Articles published in the past 12 months 4 thru 13 of 397
Parkinson"s disease over the last 100 years.
Potential nondopaminergic drugs for Parki...
Preclinical versus clinical neuroprotection.
Position of COMT inhibition in the treatmen...
Sleep attacks--facts and fiction: a critical r...
Is the cause of Parkinson"s disease enviro...
Strategies to modify levodopa treatment.
Sleep disorders in Parkinson"s disease.
Single-photon emission tomograp
Renaissance of amantadine in the
Recent Articles
Authors
74 thru 83 of 400
Corkin S
Cory-Slechta DA
Cote LJ
Cummings JL
Da Prada M
Dagher A
Dahlström A
Damecour CL
Daniel SE
Davidson MC
Experts
Organizations that published articles
9 thru 18 of 400
Cambridge Centre for Brain Repair, Univer...
Cellular and Clinical Neurobiology Program...
Center for Brain and Cognition, Department ...
Center for Gene Therapy, Tulane University...
Center for Materials of Brain Diseases, Niig...
Center for Molecular and Behavioral Neuro...
Center for Molecular and Behavioral Neur...
Center for Neurodegenerative Disease Res...
Center for Neurologic Diseases, Brigham a...
Center for Neuroscience, North Shore Univ...
Organizations
Parent diseases/phenotypes
isease/Phenotype
Pathwa
1 thru 8 of 8
Neurodegenerative Diseases
Parkinsonian Disorders
Related Diseases
Parkinson Disease
Alzheimer's disease [Homo sapiens]
Amyotrophic lateral sclerosis (ALS) [Homo ...
Huntington's disease [Homo sapiens]
Oxidative phosphorylation [Homo sapiens]
Parkinson's disease [Homo sapiens]
Prion disease [Homo sapiens]
Proteasome [Homo sapiens]
Ribosome [Homo sapiens]
Metabolic Pathways
Genes
18 thru 27 of 53
Ly [Homo sapiens]
MAPT [Homo sapiens]
MB [Homo sapiens]
MLLT7 [Homo sapiens]
MTBT1 [Homo sapiens]
MTND1 [Homo sapiens]
ND1 [Homo sapiens]
Ndufv2 [Homo sapiens]
NONO [Homo sapiens]
PARK2 [Homo sapiens]
Genes
Proteins
1 thru 10 of 400
14-3-3 protein tau [Homo sapiens]
2-5 oligoadenylate synthetase 3 [Homo sapi...
oligoadenylate synthetase 2 [Homo sa...
oligoadenylate synthetase 3 [Homo sa...
kDa subunit of complex I [Homo sapiens]
5-hydroxytryptamine 6 receptor [Homo sapi...
54 kDa nuclear RNA- and DNA-binding prot...
Aconitate hydratase, mitochondrial precur...
AD-015 protein [Homo sapiens]
[Homo sapiens]
Proteins
Protein families
Adrenergi
Alpha-synuclein
Amine oxidase
Amine oxidase (flavin-containing)
Aminoglycoside phosphotransferase
Aromatic-L-amino-acid decarboxylase
Arylamine N-acetyltransferase
Arylamine N-acetyltransferase
Aryldialkylphos
Arylester
Protein Families

Semagix Freedom Architecture

(a platform for building ontology-driven information system)

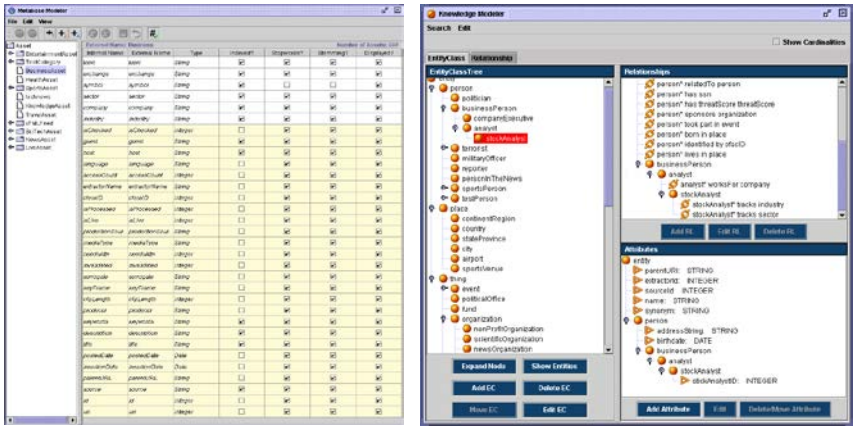


Semagix

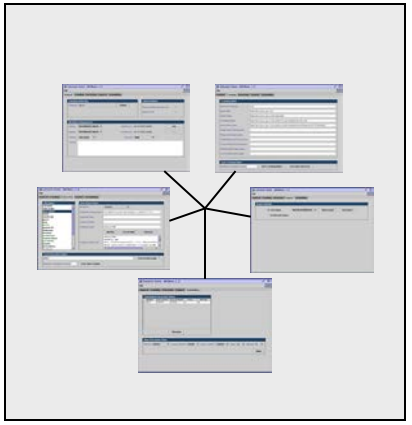
Ontologies Semagix has designed:

- Few classes to many tens of classes and relationships (types); very small number of designers/knowledge experts; descriptional component (schema) designed with GUI
- Hundreds of thousands to over 10 million entities and relationships (instances/assertions/description base)
- Few to tens of knowledge sources; populated mostly automatically by knowledge extractors
- Primary scientific challenges faced: entity ambiguity resolution and data cleanup
- Total effort: few person weeks
- Key requirement: trusted knowledge sources

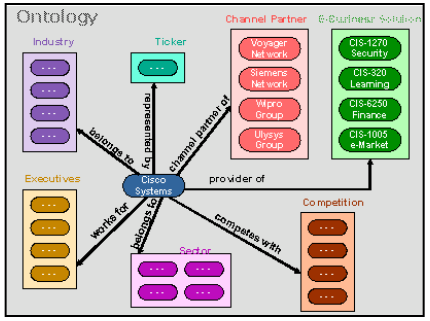
Ontology Creation and Maintenance Steps



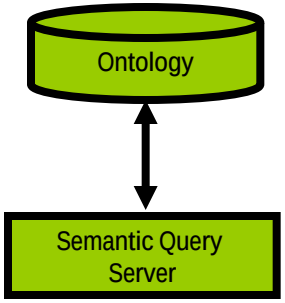
1. Ontology Model Creation (Description)



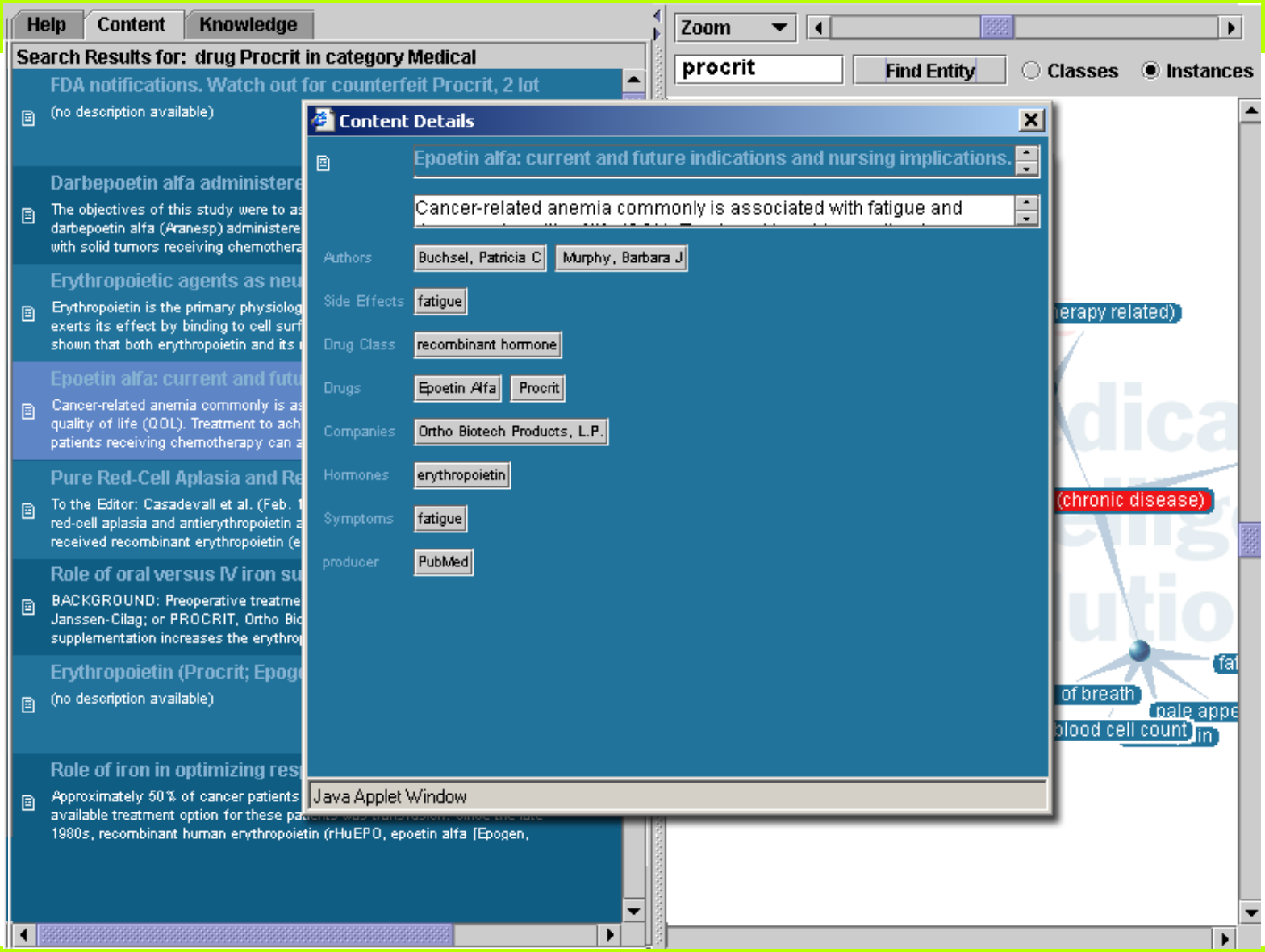
2. Knowledge Agent Creation



3. Automatic aggregation of Knowledge



4. Querying the Ontology


Help
Content
Knowledge
Search Results for: drug Procrit in category Medical
FDA notifications. Watch out for counterfeit Procrit, 2 lot
(no description available)
Darbepoetin alfa administere
The objectives of this study were to as
darbepoetin alfa (Aranesp) administere
with solid tumors receiving chemothera
Erythropoietic agents as neu
Erythropoietin is the primary physiolog
exerts its effect by binding to cell surf
shown that both erythropoietin and its
Epoetin alfa: current and futu
Cancer-related anemia commonly is as
quality of life (QOL). Treatment to ach
patients receiving chemotherapy can a
Pure Red-Cell Aplasia and Re
To the Editor: Casadevall et al. (Feb. 1
red-cell aplasia and antierythropoietin a
received recombinant erythropoietin (e
Role of oral versus IV iron su
BACKGROUND: Preoperative treatme
Janssen-Cilag; or PROCRIT, Ortho Bio
supplementation increases the erythro
Erythropoietin (Procrit; Epog
(no description available)
Role of iron in optimizing res
Approximately 50% of cancer patients
available treatment option for these pa
1980s, recombinant human erythropoietin (rHuEPO, epoetin alfa [Epogen,
Zoom
procrit
Find Entity
Classes
Instances
Content Details
Epoetin alfa: current and future indications and nursing implications.
Cancer-related anemia commonly is associated with fatigue and
Authors
Buchsel, Patricia C
Murphy, Barbara J
Side Effects
fatigue
Drug Class
recombinant hormone
Drugs
Epoetin Alfa
Procrit
Companies
Ortho Biotech Products, L.P.
Hormones
erythropoietin
Symptoms
fatigue
producer
PubMed
Java Applet Window
erapy related)
(chronic disease)
of breath
pale appe
lood cell count
in

bio ontologies

bioinformatics

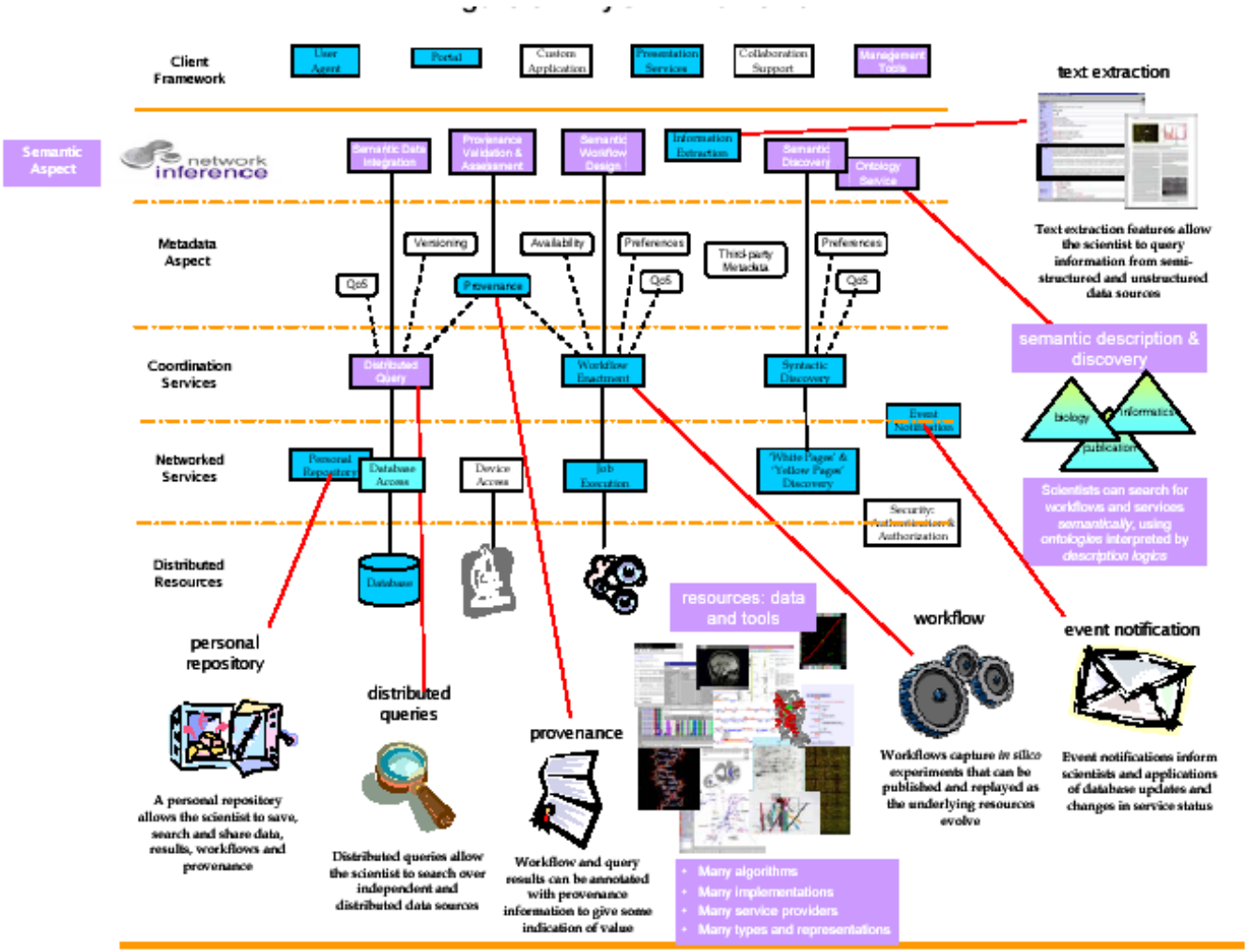
drug discovery

LSDIS

semantic web process

glycomics

Cerebra's myGrid Framework



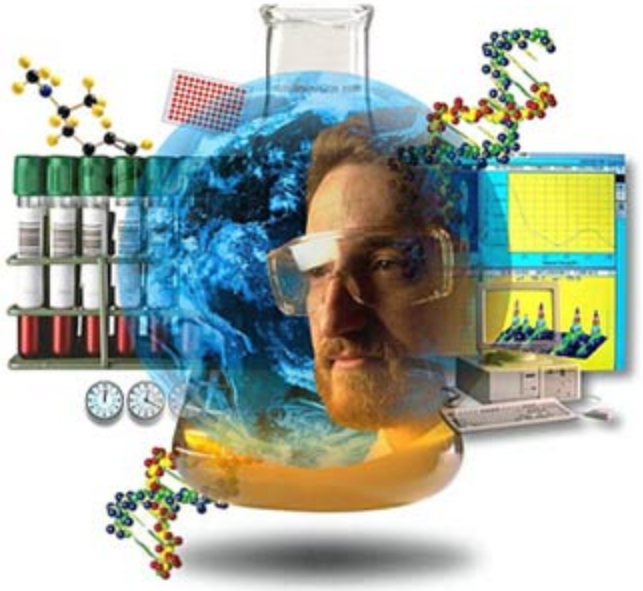
Outline

- Evolution of Science
- Challenges in biology
- What can bioInformatics do for Biology?
- What can Semantics do for BioInformatics?
  - Some examples of Semantics-powered Bioinformatics

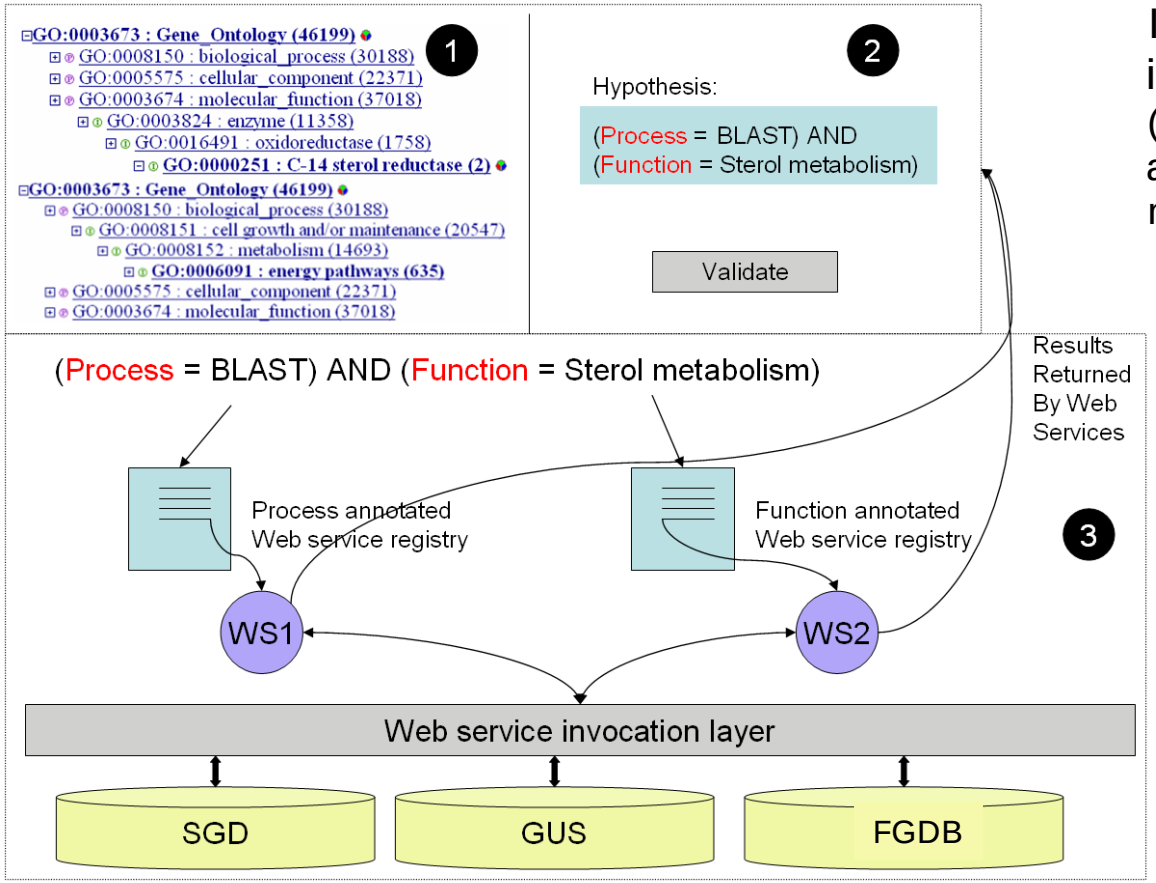
bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

Applying Semantics to BioInformatics : Example 1

Semantic Browsing, Querying and Integration



Semantic Querying, Browsing, Integration to find potential antifungal drug targets



Is this or similar gene in other organism? (most Antifungals are associated with Sterol mechanism)

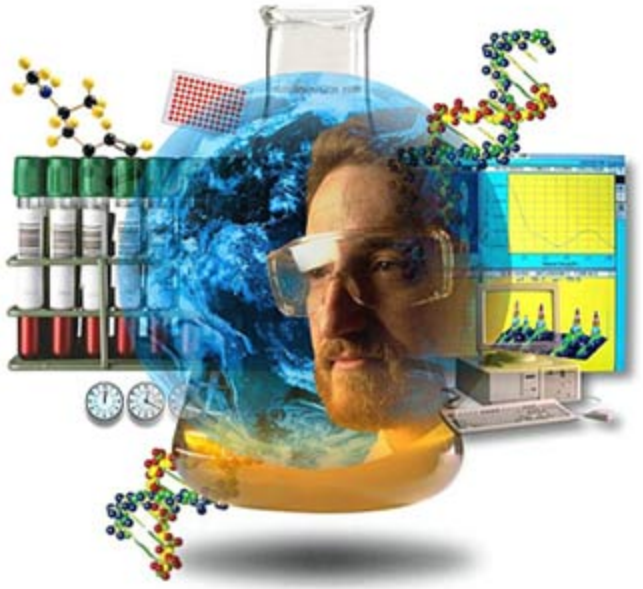
Services: BLAST, Co-expression analysis, Phylogeny; If annotated, directly access DB, else use BLAST to normalize

Databases for different organisms

bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

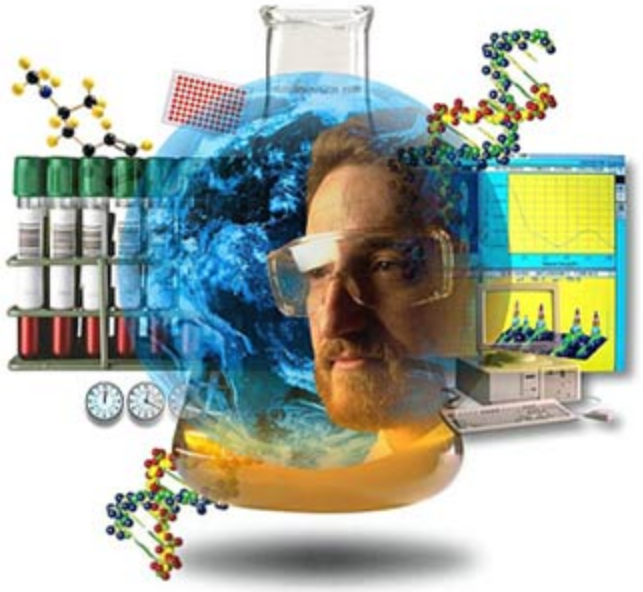
Applying Semantics to BioInformatics : Example 2

Analytics in Drug Discovery

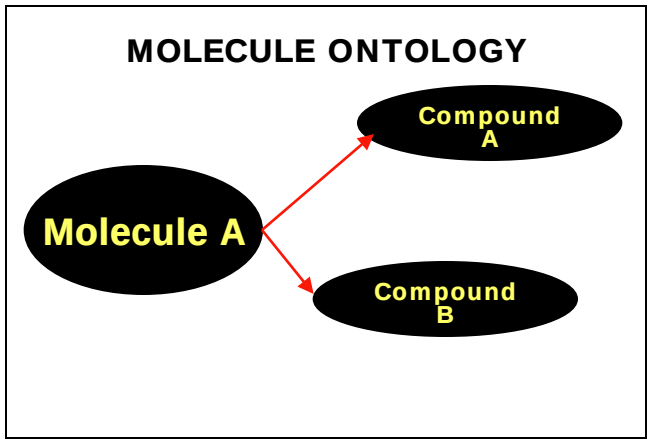


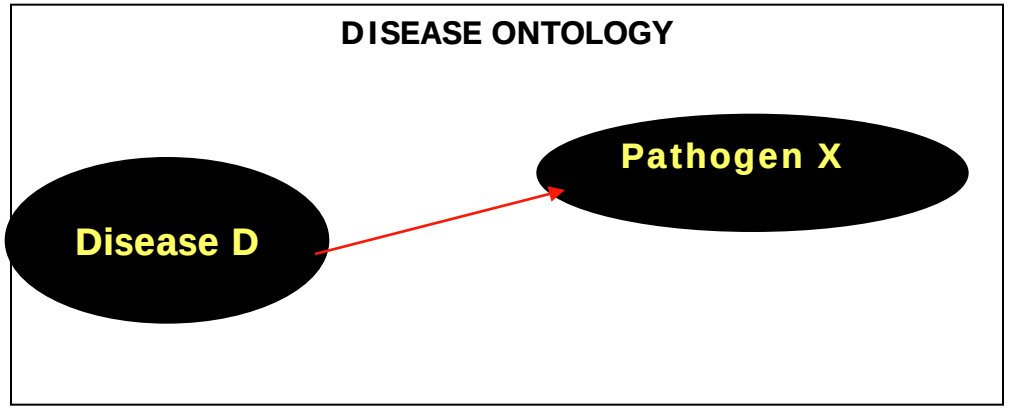
Analytics, Using Explicit and Implicit Relationships in Drug Discovery

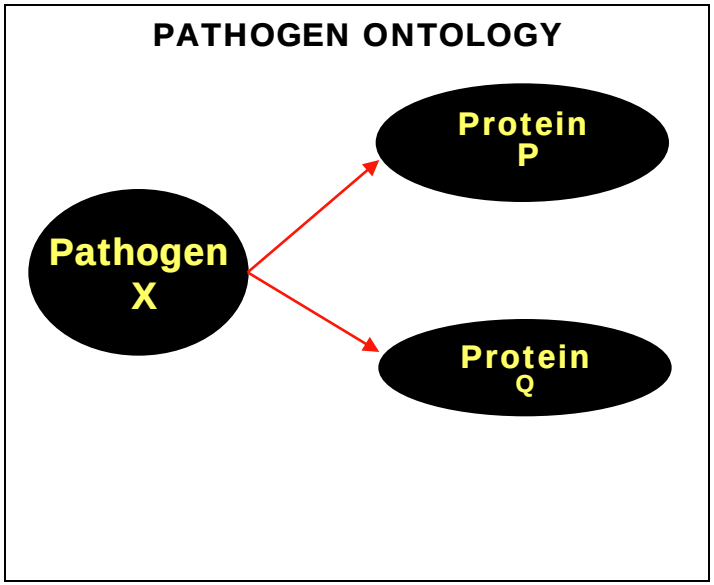
- Some molecules contain functional groups that inhibit them from acting as drugs
- Elimination of these groups can make the molecule, a DRUG.



Step 1: Capture domains using ontologies







bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

Step 2: Traverse explicit relationships

STEP 3 (Molecule Ontology):

1. Look up the molecule ontology
2. Identify the composition of the possible drug.

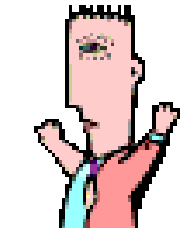
STEP 1 (Disease Ontology):

1. Look up the disease ontology
2. Identify the disease causing pathogen.

STEP 2 (Pathogen Ontology):

1. Look up the pathogen ontology
2. Identify the molecular composition of the pathogen.

Step 3: Discovering Implicit relationships…



Extract the relationships amongst the compounds of the potential drug and the pathogen.

Compound A inhibits protein P

Compound B inhibits protein Q

Compound A and B together inhibit protein P by killing protein Q

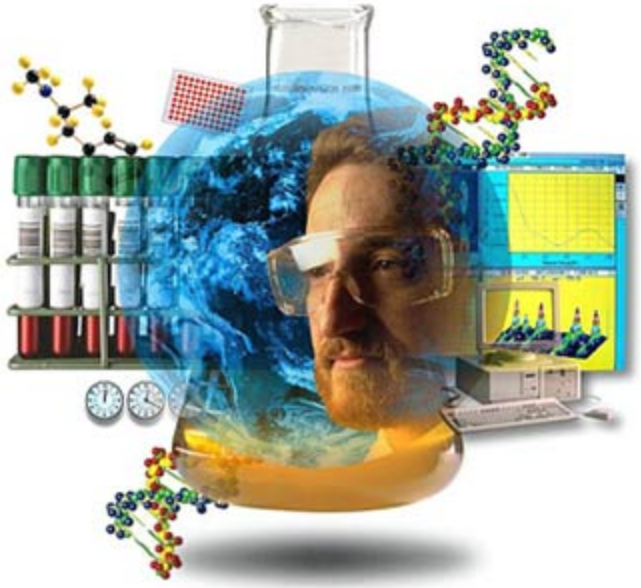
Inferences Based on Relationships

- Compound B doesn't contribute to the curing aspect of the drug, but rather generates a toxin.
- Eliminate compound B and molecule A can be a potential drug.
- However if the host has protein P, then we cannot use protein to bind the drug.
- So look for another drug that can bind at protein Q without producing a toxin
- Eliminate and Discover!!!!

bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

Applying Semantics to BioInformatics : Example 3

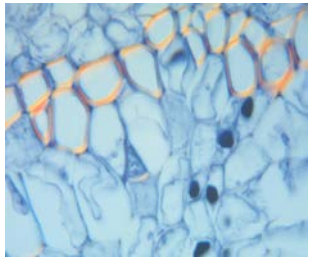
Using Ontologies in cancer research



bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

Disparate Data from Different Experiments

Experiment 1



Metastatic cancer cells

Increased GNT-V Activity

Experiment 2

Cancer marker glycan sequence elevated in glycoprotein beta 1 integrin

Knowledge Stored in Ontologies

- GO Ontology
  - GNT V is an enzyme involved in production of N-glycans
- Glycan Structure Ontology (GlycO)
  - Sequences and structures of Glycans
- Extended Structure Ontology
  - How the structures are made
    - E.g GNT V produces certain class of N-Glycans

bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

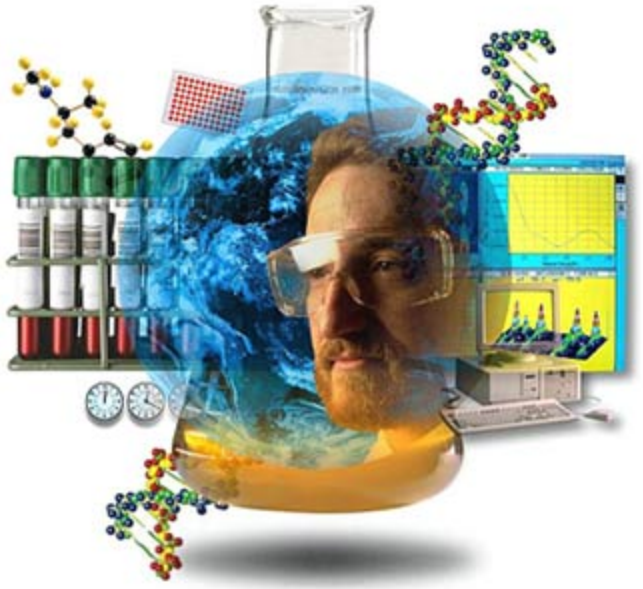
Finding New Information

- Combine data from experiments and knowledge stored in ontologies
- Known assertion from Experiments
  - Beta Integrin is involved in cancer
- New assertion to be tested
  - Are any other glycoproteins involved in cancer ?

bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

Applying Semantics to BioInformatics : Example 4

Applying Semantics to BioInformatics Processes



bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

Creating BioSemantic Processes

- **Question: *What essential genes do we all have in common?***
- Research process for this using current techniques takes long time (2 years)
  G. Strobel and J Arnold. Essential Eukaryotic Core, Evolution (to appear, 2003)
- Let us demonstrate use of Functional, Data, QoS and Execution Semantics to automate the process and reduce time

bio ontologies bioinformatics LSDIS drug discovery semantic web process glycomics

Creating BioSemantic Processes

Process for the question:

1. Input GO Id of relevant gene
2. Retrieve similar sequences
3. Perform multiple sequence alignment
4. Construct phylogenetic tree

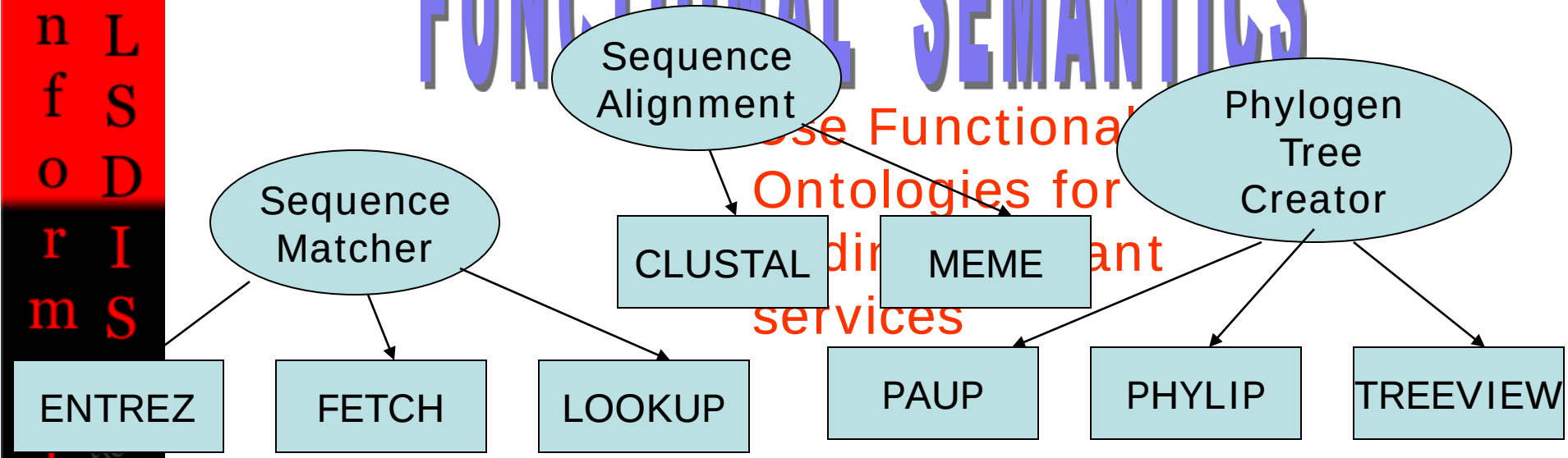
BioSemantic Process Definition

- Create semantically annotated Web Services (wrapping tools) and process using METEOR-S
- Semantic templates at each node allow choosing multiple tools automatically
- Run multiple instances of process using different tools, combine results

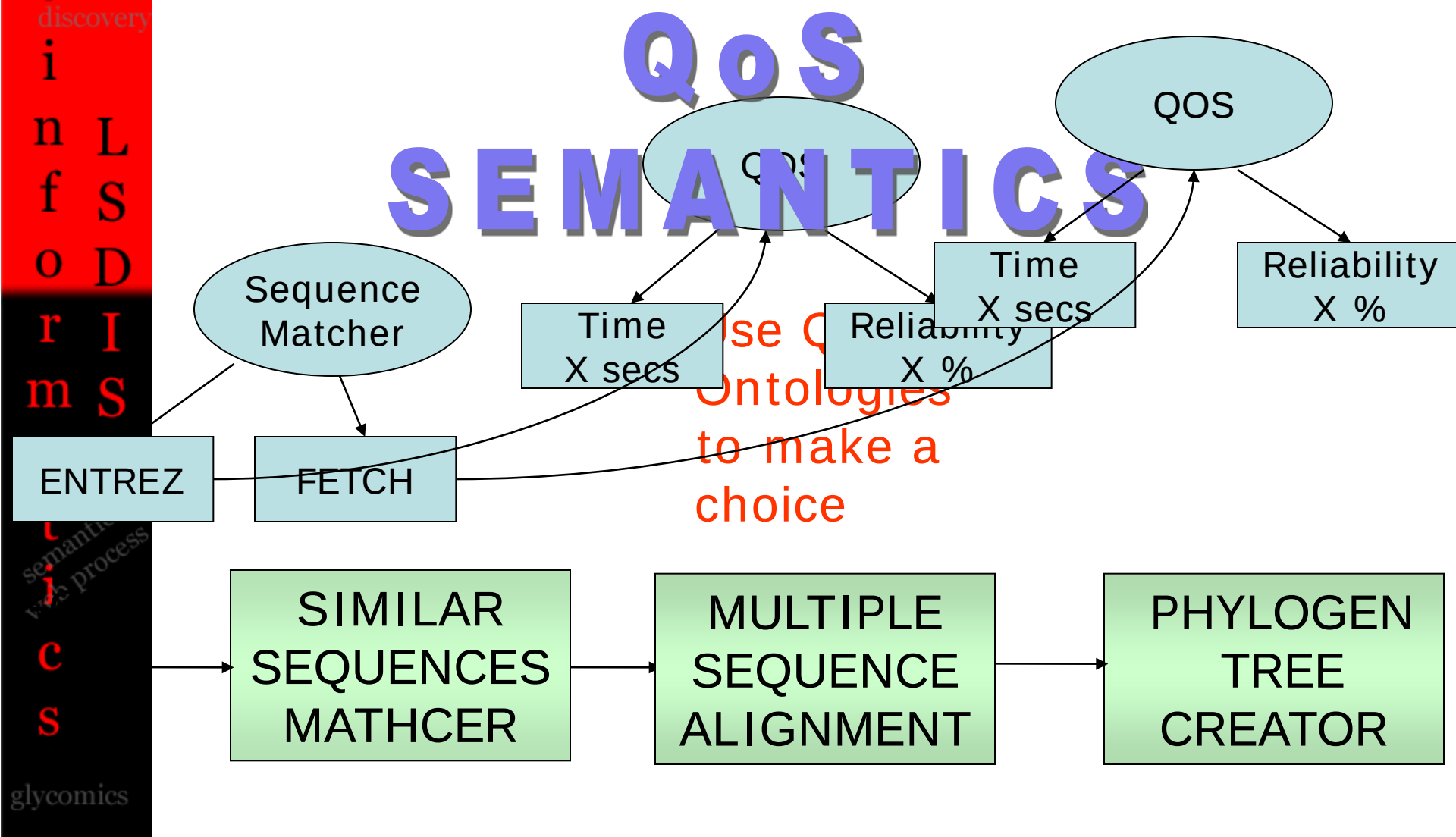
Semantic Bioinformatics Processes

FUNCTIONAL SEMANTICS

Use Functional Ontologies for discovery relevant services



Semantic Bioinformatics Processes



bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

Semantic Bioinformatics Processes

DATA SEMANTICS

MAY REQUIRE DATA CONVERSION

MAY REQUIRE DATA CONVERSION

Use Concept Ontologies for Interoperability

SIMILAR SEQUENCES MATHCER

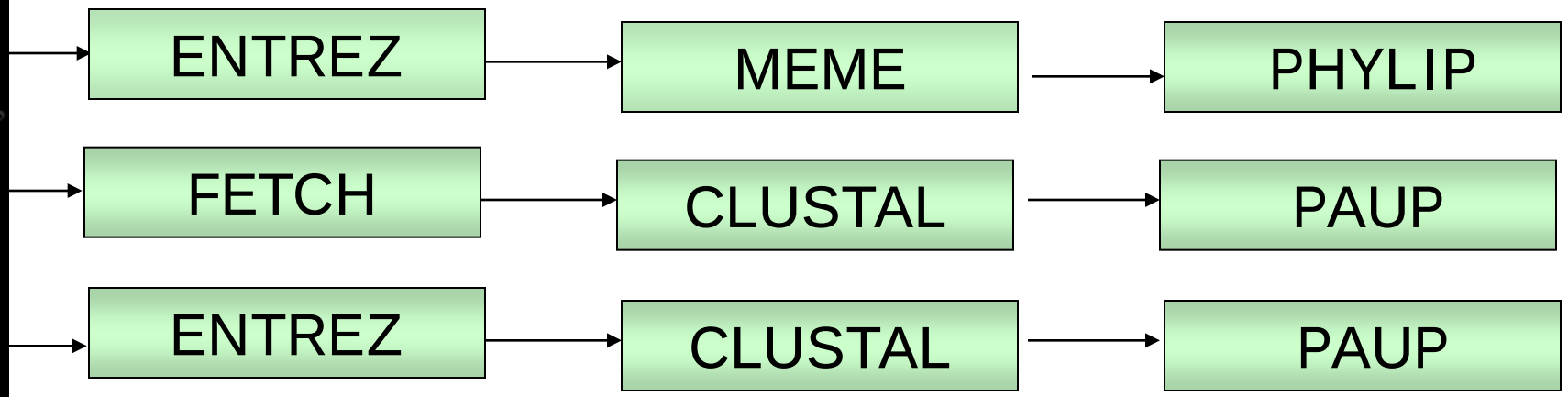
MULTIPLE SEQUENCE ALIGNMENT

PHYLOGEN TREE CREATOR

Semantic Bioinformatics Processes

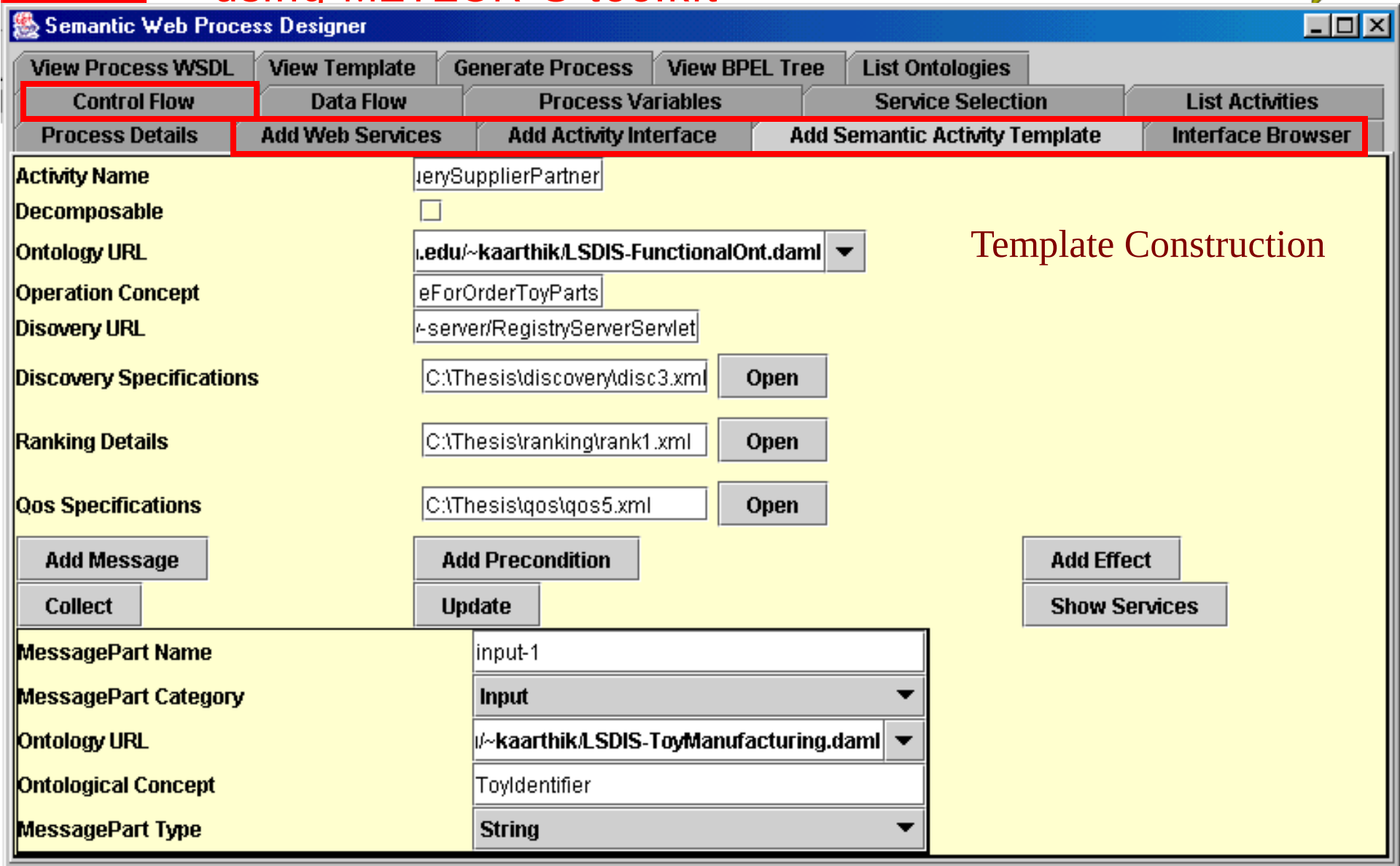
EXECUTION SEMANTICS

Use Execution Semantics for execution monitoring of different instances



Semantic Web Process Mgmt

using METEOR-S toolkit



Template Construction

bio ontologies

bioinformatics

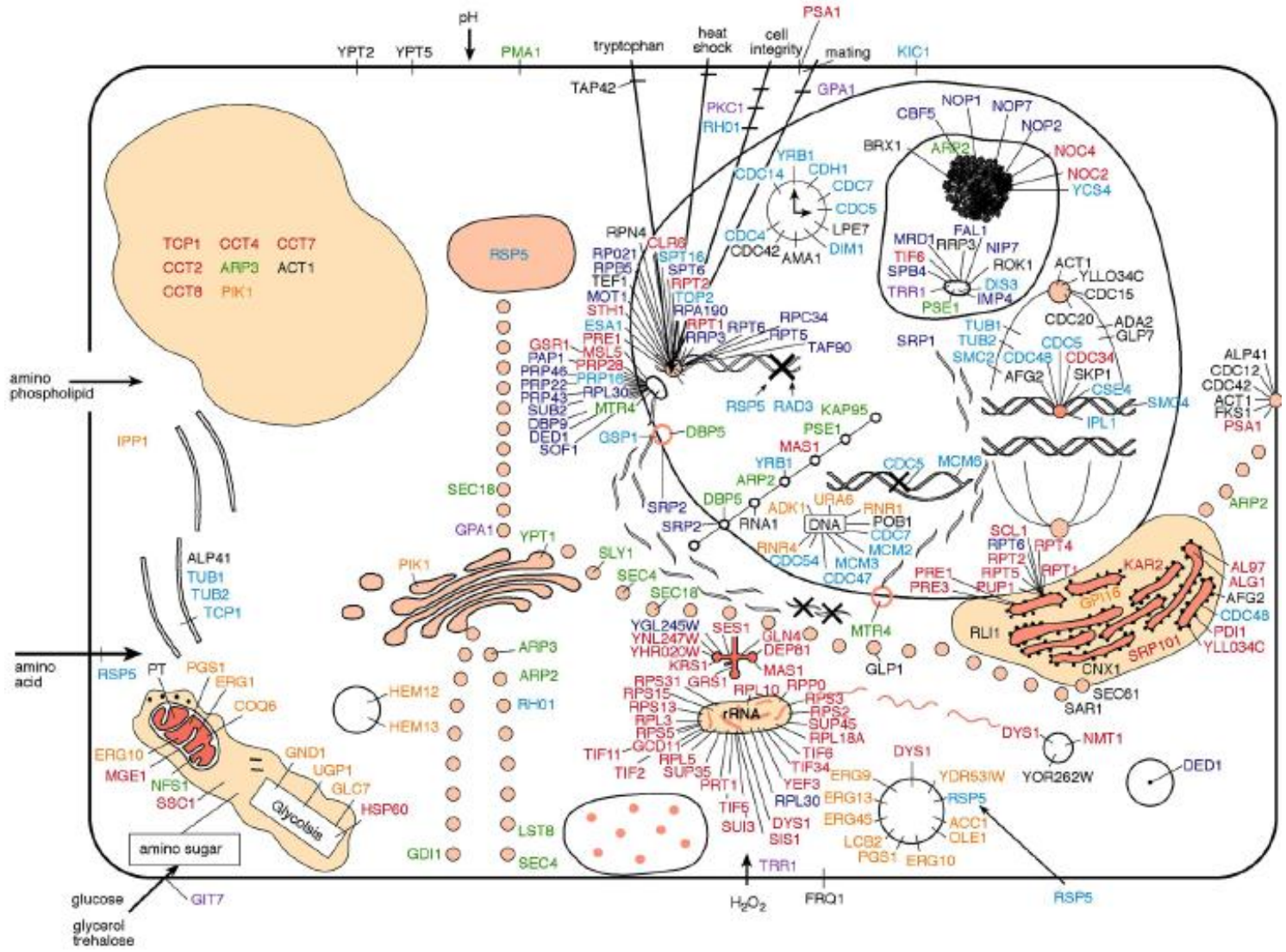
drug discovery

semantic web process

glycomics

LSDIS

Common genes



bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

Conclusion

- Biology research at unique standpoint in history
- Earlier
  - Biology divided into many sub-fields
  - Significant progress in those domains
- Present
  - Searching for the bigger picture
  - Need to combine knowledge from sub-fields
    - Disparate sources, terminologies
- Semantics is the vehicle to get to the answers

Conclusion

- Need to capture the bioinformatics domains using ontologies
  - Not just schema, instances are required
- Presented the use of semantics in
  - Search, Integration
  - Analytics, Knowledge discovery
  - Process Automation
- Collaboration between scientists and computer scientists with semantic techniques and tools make hard things easier

bio ontologies bioinformatics drug discovery LSDIS semantic web process glycomics

Sources

1. Picture on each sub-section title : http://www.3rdMill.com
2. Pictures on the title collage were taken from
   - http://web.mit.edu/beh.480j/www/covermed.gif - Glycomics
   - http://bcf.bcm.tmc.edu/proteomics_3.jpg - Proteomics
   - http://www.microsoft.com/spain/msdn/eventos/presentaciones/images/ - Webservices
   - http://panda.cs.inf.shizuoka.ac.jp/~cs8060/images/ontology.png - ontology
   - http://www.chem.agilent.com/cag/other/workflow-330.gif - RNA analysis workflow
   - http://www.exploratorium.edu/ti/human_body/dna.html - Genomics
   - Molecular Modelling – Principles and Applications by Andrew R Leach - Drug Discovery

Bioinformatics and Semantic Web at LSDIS Lab, UGA:
http://lsdis.cs.uga.edu/proj/glycomics/

Also, Semantic Web Processes: http://swp.semanticweb.org